人文社科领域中文通用大模型性能评测*

赵志枭[1,2]　胡蝶[1,2]　刘畅[1,2]　沈思[3]　王东波[1,2]
[1] 南京农业大学信息管理学院　南京　210095
[2] 南京农业大学人文与社会计算研究中心　南京　210095
[3] 南京理工大学经济管理学院　南京　210094

摘要：[目的/意义]本文以人文社科领域为出发点，从人文社科领域基础知识与人文社科学术文本两个方面入手进行人文社科领域模型性能比对。旨在为人文社科领域提供一份体系化的大模型评测基准，供人文社科相关领域研究人员参考。**[方法/过程]**设计了 7 个人文社科领域相关的评测任务并选取对应指标，在此基础上，选取了当前开源且性能较优的通用领域中文大模型，通过调用本地模型以问答形式完成领域化任务，并选取相关指标对其在人文社科领域的性能进行了量化评测。**[结果/结论]**评测结果表明，在本文选取的开源模型中，无论是基座模型还是对话模型，Qwen 性能最优，Baichuan2 紧随其后，InternLM 次之，Atom 表现最差，此外，大多数情况下，相较于基座模型，对话模型表现出了更加优越的性能。

关键词：人文社科　大模型评测　领域知识　学术文本
分类号：G352

1　引言/Introduction

随着越来越多的互联网公司和科研团队投身于 AIGC 的浪潮中，大量开源可商用的大语言模型被推出，社会各界都意识到了这场人工智能革命带来的机遇，纷纷加入到大模型研究中。经过了长达半个世纪的发展，人文社科与计算机科学已经形成了深度的交叉融合，人文计算、社会计算等交叉学科展现出宽广的发展前景，作为计算机与人文社科的交叉融合，其旨在将计算机科学的思想理论应用于人文社科各个领域，丰富人文社科研究内容[1,2]。伴随着人工智能技术的不断推进，人文社科与计算机科学的交叉融合也迈入了新的阶段，人文社科领域化的数据、理论与人工智能技术相结合，必将带来两个领域的协同发展。

由于大模型对于硬件设备、数据的高需求，当前绝大多数科研团队很难完成从零开始的通用领域大模型训练，而随着大模型研究的发展，模型训练流程逐渐趋同，开源大模型的推出让更多普通科研团队可以结合特定领域构建垂直大模型。垂直领域模型的构建，往往使用领域数据对已有模型进行增量训练，使模型在保持强大语言能力的同时学习更多垂直领域专业知识，由此可见，基座模型的选取对于最终模型性能有着极大的影响。大模型的发展也引发了大模型评测的浪潮，其中也包含针对垂直领域的评测，而目前面向人文社科领域的大语言模型评测较为匮乏。对于大部分人文社科领域学者来说，从大量开源通用大模型中选取适合自己研究领域的大语言模型较为困难，无论是大语言模型在人文社科领域的应用还是人文

---

*本文系国家社科基金重大项目“中国古代典籍跨语言知识库构建及应用研究”（项目编号：21&ZD331）研究成果之一。
This work is supported by Social Science Foundation of China project “Research on the construction and application of cross-language support library for ancient Chinese classics” (Grant No. 21&ZD331).

作者简介：赵志枭，硕士研究生；胡蝶，硕士研究生；刘畅，博士研究生；沈思，博士，副教授，博士生导师；王东波，博士，教授，博士生导师，通信作者，E-mail：db.wang@njau.edu.cn。
Author: Zhao Zhixiao, Master student; Hu Die, Master student; Liu Chang, Doctoral student; Shen Si, Associate professor, PhD, Doctoral supervisor; Wang Dongbo, Professor, PhD, Doctoral supervisor, Corresponding author, E-mail: db.wang@njau.edu.cn.

学者对新时代新技术的应用，完备的评测体系必不可少。因此，本文针对人文社科这一较为宽泛但有其内在特征的领域开展中文开源大模型评测，旨在为人文社科领域提供一套用于评测大模型的基准，供相关领域研究人员参考。对于无计算机基础的人文学者来说，本文可以为其了解、使用大语言模型提供量化参考，而对于计算人文研究人员来说，也可以通过本文的评测结果选择适合的大语言模型开展人文社科领域的大模型研究。

2 相关研究/Related research

2.1 人文社科与人工智能的交叉融合

数字人文这一新兴学科，在信息资源管理专业掀起了研究热潮。黄水清等[3]对计算人文当前的学科发展进行了梳理，以计算文学、计算语言学、计算史学等为例论述了“计算 X”学科的发展状况。计算机技术与众多人文学科相融合，数据驱动的研究范式在人文学科中广泛应用，产生了一系列以人文学科为基础的实证性研究。王东波等[4]基于四库全书数据和 BERT 模型，构建了面向古文信息处理的领域化预训练模型，在众多古籍信息处理任务中表现优越。张卫等[5]将古诗文本鉴赏与人工智能技术相融合，将古诗文字向量与预训练模型相结合，实现古诗文本情感术语的自动高效抽取。张琪等[6]以《史记》为对象，将史书中的复杂信息进行知识重组和形式再造，构建了以史书多维知识库和知识可视化平台，消减了用户阅读史书的障碍。喻雪寒等[7]将神经网络与机器阅读理解模式相融合，分别在编年体史书和纪传体史书语料上进行了训练和验证。在数字人文领域，古籍智能处理、中华传统文化挖掘已经成为非常热门的研究内容，以人工智能技术推动古籍活化利用成为了人文学科与人工智能技术结合的典范。除此之外，金融、法学等社会科学与人工智能的交叉融合也越发紧密，张瑞祥等[8]以计算法学学科发展路径为底层逻辑，结合人工智能技术在法律领域的应用论述了计算法学研究范式的变迁和现状。梁祝等[9]利用类新闻事实文本，结合法律判决文书的结构内容特征构建了判决文书推荐系统。

信息技术的更新迭代为人文社会科学领域注入新的活力，不断推动人文社科研究的进步与发展。当前，大语言模型技术使得人工智能实现了阶段性跨越，在大模型时代下，已有研究人员就通用大语言模型在人文社科领域的适用性进行了探究。为更好地适应人文社科领域专业化数据与个性化任务需求，产业界与学术界推出了更多面向人文社科垂直领域的大语言模型。例如，在金融领域，W. Shijie 等[10]基于大批量的金融相关数据，构建了 50B 参数量的金融领域大模型 Bloomberggpt，填补了金融领域大模型的空白。Y. Hongyang 等[11]提出了一种新的金融大模型开源框架，通过开源数据微调提升模型领域能力而降低训练成本，为金融人工智能的发展推波助澜。后续的 PIXIU[12]、InvestLM[13]等同样是大语言模型在金融领域的重要尝试。在法律领域，LawGPT_zh[14]使用 ChatGPT 清洗开源法律数据集，从而构建了适用于法律领域的开源大模型。ChatLaw[15]则基于 Ziya-13B 等基座模型，使用大量法律新闻、法律论坛、法条、法考题、判决文书等原始文本构造的对话数据进行模型构建。除此之外，教育[16,17]、医疗[18,19]、电子商务[20,21]等领域均有相应的大模型成果，为专业场景下的应用提供了可靠支持。

总的来说，人工智能技术已经应用于人文社科的各个领域。大模型时代下，垂直领域大模型的构建变得愈发重要，领域数据集的构建成为人文社科领域深入开展大模型研究的基础。特定领域数据集构建，必然需要特定领域研究人员的参与，而数据的整理、存储流程是信息资源管理的重要研究内容，大模型的训练、微调又涉及计算机科学的相关理论技术，大模型构建必然涉及到跨专业、跨领域的深度合作与交流。

2.2 大模型评测相关研究

当前大模型评测主要可以分为几个方面：语言能力，大语言模型最令人称道的就是其强大的语言能力，在少样本甚至零样本情况下即可胜任大多数自然语言处理任务；知识储备，由于预训练阶段学习的大量数据，大模型拥有强大的知识储备，可以应对一些事实问答，例

如金融、法律、医疗等领域相关问答和一些常识问答；安全性，2023 年 7 月发布的生成式人工智能服务管理暂行办法中明确规定，提供、使用生成式人工智能，应当遵守法律、行政法规，尊重社会公德和伦理道德[22]；其他内容，包括大模型在实际生活中的应用，将大模型作为智能代理工具等。

模型语言能力评测可以分为自然语言理解和自然语言生成两类任务，自然语言理解方面，主要包含情感分析、文本分类等任务，自然语言生成方面，主要包含对话、摘要、翻译等任务。大模型语言理解能力评测方面，W. Zengzhi 等[23]对 ChatGPT 理解文本中包含观点、情绪、情感的理解能力进行了评测，并与微调后的 BERT 模型以及最优方法进行比对以判断 ChatGPT 能否胜任情感分析任务。Z. Wenxuan 等[24]对大模型在各类情感分析任务上的性能进行了调查，并将大模型与在特定数据集上训练后的小模型进行对比，结果显示，大模型在简单任务中表现良好，在较复杂任务中表现较差，但只要增加少量样本，仍然表现出了优于小模型的性能。P. Alejandro 等[25]评测了大模型在公共事务文件分类方面的性能，构建了包含 30 个类别的文本分类数据集，并为每个类别数据构建了二分类评测数据，通过二分类评测的方法解决了类别样本不平衡问题。大模型语言生成能力评测方面，Z. Wenhao 等[26]针对大语言模型的翻译能力开展研究，包括对大模型翻译能力的评测、大模型翻译能力激发和大模型在不同语言上翻译能力的表现，探讨了大模型翻译能力提升的路径及训练语料语种对于大模型翻译能力提升的影响。

在知识储备方面，有针对垂直领域的领域知识评测，也有针对通用领域的世界知识评测，例如 D. Xuanquy 等[27]对 ChatGPT 在回答高中数学多项选择题时的表现进行了分析，结果显示，ChatGPT 很难回答导数、空间几何等方面的问题，而在指数、对数等问题上表现出色。W. Yiran 等[28]探索了 GPT-4 解决高难度数学问题的能力，值得注意的是，该研究使用了多种调用 GPT-4 的方式进行数学测试，其中包括 MathChat。D. Dat 等[29]评估了大模型在遗传学领域的表现，ChatGPT 的表现与人类相近，在记忆性问题方面表现尤其突出。G. Aidan 等[30]评价了 ChatGPT 在医疗护照考试中的表现，使用 2 组多项选择题对 ChaGPT 进行评估，并与另外两个大模型 GPT-3、InstructGPT 进行了比较。

在安全性方面，主要针对大模型道德偏见、鲁棒性等方面开展评测，Z. Jiaxu 等[31]构建了用于评价中文对话大模型的数据集 CHBias，使用该数据集对对话大模型进行了评测，证明了一些模型仍然存在社会偏见倾向。P. Alicia 等[32]引入了问答偏见基准 BBQ，涉及美国社会环境中九种即时存在的社会偏见，评测结果发现，给定上下文能一定程度上消减模型的刻板映像，但不能完全消除。

除此之外，还有大量跨领域跨任务的大模型评测基准被提出，例如 Z. Wenxuan 等[33]采用 9 种不同语言构建的来源于真实试题的评测数据集 M3Exam，H. Yuzhen 等[34]构建了包含 4 个难度级别、52 中学科来源的多项选择评测数据集，X. Liang 等[35]结合大模型发布平台用户评分对大语言模型进行了评价，探讨了封闭式问题评价大语言模型的缺陷。

综上所述，当前大语言模型评测很多都是基于 ChatGPT 开展，主要原因在于 ChatGPT 已经成为当前大模型的标杆，其强大的指令跟随能力也便于人们进行评测，也有研究使用 ChatGPT 对其他大模型输出内容进行评测，这对于一些参数量较小的大模型来说不失为一种可行的方法。另一方面，很多大模型评测基准采用了大量选择题作为测试任务，使用选择题进行评测为评测结果量化提供了便利，但是，封闭式的任务或许并不能完全体现模型的性能，与之相对的，开放式任务在量化评测方面很难同时顾及准确、客观、便捷。

3　评测体系设计/Evaluation system design

在近期推出的大模型评测任务中，单项选择题成为了评测的重要组成部分，究其缘由，通过单项选择对模型进行评测，可以相对客观、快捷地获取模型在某个方面的得分。但是，单项选择题着重考察模型对特定领域的知识储备和指令跟随能力，模型的文本生成能力一定

程度上被忽略了。为尽可能全面地对大模型在人文社科领域的表现进行评测，本文从领域知识与学术文本两个角度入手，挑选了 13 个性能较为优越的通用领域开源中文大模型，设计了 7 个任务开展评测。本文的整体框架结构如图 1 所示。

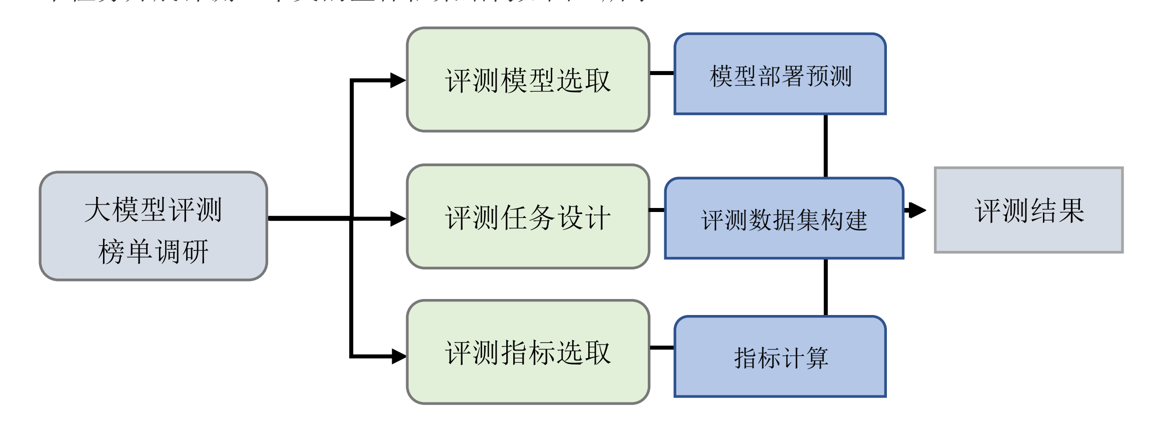


图 1 大模型评测整体流程框架

Figure 1 Framework for Large Model Evaluation

3.1 评测模型选取

在评测模型选取初期，调研了当前认可度较高的大模型评测榜单，例如 OpenLLM[36]、SuperCLUE[35]、C-eval[34]、CLiB[37]等，选取了其中性能较为优越的开源中文大模型作为本文模型的候选模型。根据调研结果可以发现，对于同一模型，参数量的变化会大幅影响模型性能，而当前大多数开源中文模型参数量均集中在 7B 左右，因此，本文同一选取参数量在十亿量级的大模型进行评测，具体模型信息如表 1 所示。本文共选取了 13 个模型，其中包括基座模型 6 个和对话模型 7 个，对于其中一些模型，例如 ChatGLM-6B 模型，当前仅开源了对话模型，因此只选择了对话模型版本。相较于基座模型，对话模型由于经过通用对话数据微调，有着更加强大的指令跟随能力，能够更好地理解用户提问并应答，以当前大多通用领域模型评测的结果来看，对于同一模型来说，在以对话为任务形式的评测下，对话模型往往可以表现出更加优越的性能。因此，本文做出假设：在垂直领域，对话模型仍然可以表现出更加优越的性能，在评测过程中，将对话模型与基座模型分别进行对比，使模型性能的对比更加清晰。在评测任务的形式上，由于基座模型指令跟随能力较弱，因此，对于一些格式化要求较高的任务，着重对基座模型的输出进行了后处理，例如单项选择、分类等任务。

表 1 中文开源大模型选取

Table 1 Chinese Open Source Large Models

模型名称	模型类型	发布机构
Atom-7B[38]	基座模型	FlagAlpha
Baichuan-7B[39]		百川智能
Baichuan2-7B[40]		
Chinese-Alpaca-7B[41]		ymcui
InternLM-7B[42]		上海人工智能实验室
Qwen-7B[43]		阿里云
Atom-7B-Chat[38]	对话模型	FlagAlpha
Baichuan2-7B-Chat[40]		百川智能
ChatGLM-6B[44]		智谱 AI
ChatGLM2-6B[45]		
InternLM-7B-Chat[42]		上海人工智能实验室

Phoenix-Inst-Chat-7B[46]	香港中文大学（深圳）
Qwen-7B-Chat[43]	阿里云

3.2 评测任务构建

针对人文社科领域知识，本文构建了 3 个任务进行评测，分别是单项选择、术语解释、开放问答。单项选择、开放问答数据来源于人文社科领域基础知识竞赛试题，包括近三年内的大学生人文知识竞赛、哲学、法律、经济学领域知识竞赛试题，为保证评测涵盖学科范围更加广泛、均衡，本文按照 4:2:2:2 的比例从各类试题中抽取数据。其中，单项选择题主要包括哲学、历史、汉语文化常识、法律、经济学基础知识，开放问答以人文社科领域相关论述题为主，包括对于某一专业知识的解释、对领域化世界知识的列举，考察大语言模型对领域化知识的掌握以及分点作答能力，例如，“当代人文科学呈现应用化趋势，表现在哪几个方面”。术语解释数据来源于南京大学外国语学院术语与翻译跨学科研究基地研究成果[47]，包含了管理学、教育学、经济学、考古学等十个学科的术语文本及解释数据，从各学科数据抽取了二十条解释较为全面、完整的数据共 200 条数据作为术语解释测试语料。

针对人文社科学术文本，本文构建了 4 个任务进行评测，分别是论文摘要文本学科分类、论文摘要文本语步识别、论文标题生成和学术文本翻译。论文摘要、标题及其对应学科和语步划分来源于 CSSCI 收录期刊论文摘要文本及其对应学科分类，考虑到 CSSCI 当前共划分了 26 个学科分类，过多的选择不利于大模型输出正确的内容，因此，本文从 10 个学科的数据中，每学科每个任务抽取 20 条数据共 200 条数据作为摘要文本分类、语步识别和论文标题生成的测试语料。文本翻译数据与术语解释数据来源相同经过同样的处理流程获得，考虑到当前中文大模型均具备双语能力，文本翻译任务通过中英互译的方式对模型进行更加全面的评测。

评测所需数据收集完成，需要针对不同的任务构建提示指令，构建提示指令需要结合相关任务和期望输出进行，以单项选择为例，为降低后续评测指标计算难度，需要保证模型尽可能只输出正确答案对应选项，如果只是输入问题及相关选项，大部分情况下会输出一些无关内容，不便于后续对模型输出结果进行评测。另一方面，在提示过程中，为模型设定角色，也可以提升模型的输出效果[48]，考虑到本文的领域性，在构建提示指令时也添加了相关的角色带入用语。之后，选取少量数据在各模型上进行测试，尽可能使大多数模型能够输出符合要求的内容，构建相应指令如表 2 所示。

表 2 提示指令示例

Table 2 Examples of prompts

任务	提示指令
开放问答	作为一名人文社科领域研究人员，回答下面给出的人文社科领域开放性问答，要尽可能全面成体系 {input}
单项选择	作为一个人文社科领域研究人员，回答下面给出的单项选择题，只需要输出选项，即 A、B、C、D 中的一个，不需要解释，不要输出其他内容 {input}
术语解释	作为一名人文社科领域研究人员，对下面给出的领域术语进行解释，解释要尽可能详细 {input}

标题生成	根据下面给出的人文社科领域摘要文本，给出最有可能的标题，只需要输出标题，不要输出其他内容 {input}
文本翻译	将下面这段{文化学}领域英文（中文）文本翻译为中文（英文），只需要输出翻译结果 {input}
学科分类	根据下面给出的人文社科领域摘要文本，判断该摘要属于人文社科领域的哪个专业，从['哲学', '历史学', '法学', '政治学', '经济学', '社会学', '教育学', '心理学', '管理学', '语言学']中选择一个类别进行输出，只需要输出类别名称，不要输出其他内容 {input}
语步识别	根据下面给出的人文社科领域论文摘要文本，判断这段文本是属于摘要中的哪个部分，从['结果','方法','目的','局限','结论']中选择一个类别进行输出，只需要输出类别名称，不要输出其他内容 {input}

经过测试，完成基础指令的构建之后，需要确定的是模型提示模式。大模型强大的文本生成能力一定程度上来源于其强大的上下文学习能力，因此，在某些任务下，为尽可能提高模型输出内容的质量，可以给定少量示例以便模型可以更好地理解指令，输出更加符合需求的内容，即 few-shot[49] 。但是，并非所有任务、模型都适合 few-shot 模式，在构建指令的基础上，本文针对不同任务试验了 0-shot、1-shot、3-shot 模型下不同模型的输出质量。在尽可能保证模型能够输出理想内容的情况下降低提示示例个数，最终确定了各任务示例个数如表 3 所示。

表 3 评测提示模式

Table 3 mode of prompts

任务	提示模式
开放问答	0-shot
单项选择	0-shot
术语解释	0-shot
标题生成	0-shot
文本翻译	0-shot
学科分类	3-shot
语步识别	1-shot

3.3 评测指标选取

生成式模型的发展为模型评测带来了极大的挑战，在当前自然语言处理评价指标的基础上，本文针对各个特定任务选取了对应指标，以期尽可能地对模型生成内容进行客观、定量的评价，本文采取的评价指标及计算方法如下文所述。

(1) 准确率(Accuracy)

准确率计算简单、便捷，适用于单项选择和分类任务，但对于模型输出内容的规范性要求较高，在本文过程中，不乏有模型由于输出内容不规范导致得分较低。在单项选择、分类任务中，获得各模型输出结果之后，通过人工对结果进行了校对，将一些答案正确但输出多余内容的结果进行了修改。

(2) 精确度、召回率、调和平均值(Precision、Recall、F1-score)

精确度、召回率、调和平均值是自然语言处理任务中常用的指标，该指标的计算基于混淆矩阵进行，通过混淆矩阵，对于每一个类别，可以将全部预测结果分为四类，即预测为该类别且实际为该类别(TP)、预测非该类别且实际非该类别(TN)，预测为该类别但实际非该类别(FP)，预测非该类别但实际为该类别(FN)，得到相应的相应计算公式如下：

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 - Score = \frac{2precision * recall}{precision + recall} \tag{3}$$

考虑到生成式模型输出的不规范问题，对于多分类任务指标，采取加权平均算法得到分类任务各模型的最终得分。加权平均是使用每个类别样本数量占所有类别的样本总数的比例作为权重，再计算各指标平均值，这样可以有效屏蔽模型输出的不规范内容，加权平均计算公式如下，其中，x_i为模型所预测的 i 类别样本量占整个数据集样本量的比例，f_i为 i 类别对应的调和平均值。

$$W_x = \sum_{i=1}^{n} x_i * f_i \tag{4}$$

(3) BLEU

BLEU[50]指标是基于 n-gram 思想针对精确率制定的一系列用于衡量生成文本与参考文本之间相似度的指标。本文采用了 BLEU 作为评价文本翻译的评测指标，其具体计算公式如下：

$$BLEU - n = \frac{\sum_{S\in\{Candidate\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{Candidate\}}\sum_{gram_n\in S} Count(gram_n)} \tag{5}$$

其中，n 代表 n-gram 的长度，$Count_{match}$代表生成文本与参考文本中共同出现的 n-gram 最大数量。分母则代表生成文本中 n-gram 的个数。

(4) ROUGE

ROUGE[51]指标同样基于 n-gram 思想针对召回率制定的一系列用于衡量生成文本与参考文本之间相似度的指标。本文采用了 ROUGE-N 和 ROUGE-L 作为评价标题生成、术语解释和开放问答的评测指标。ROUGE-N 具体计算公式如下：

$$ROUGE - N = \frac{\sum_{S\in\{Reference\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{Reference\}}\sum_{gram_n\in S} Count(gram_n)} \tag{6}$$

其中，n 代表 n-gram 的长度，$Count_{match}$代表生成文本与参考文本中共同出现的 n-gram 最大数量。ROUGE-L 基于 LCS(longest common subsequence)思想，通过计算两个句子的最大公共子序列对应 P、R、F 值，具体计算公式如下：

$$R_{lcs} = \frac{LCS(X,Y)}{m} \tag{7}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \tag{8}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs} * P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{9}$$

其中，m、n 分别代表参考文本与生成文本的长度，在计算过程中，β会被设置为一个很大的数值，与 ROUGE-N 相同，ROUGE-L 主要参考召回率。

(5) chrF

chrF[52]指标与 BLEU 指标类似，区别在于，chrF 指标是基于字符级进行计算，主要针对 F-score 进行计算，具体计算公式如下：

$$\mathrm{chrF}_{\beta} = (1+\beta^2)\frac{chrP * chrR}{\beta^2 * (chrR + chrP)} \tag{10}$$

其中，chrP 为生成文本中属于参考文本的字符数所占百分比，chrR 为参考文本中属于生成文本的字符数所占百分比。chrF 指标同样应用于文本翻译的评测，为保证 BLEU、chrF 指标计算的准确性，本文采用了 sacreBLEU[53]提供的方法进行计算。

(6) MAUVE

MAUVE[54]指标从 KL 散度的角度对人工智能生成文本与人类生成文本的相似性进行评估。该指标的计算依赖于自回归语言模型，根据 MAUVE 计算的相关实验，MAUVE 指标与人类评价的相近度随着采用模型参数量的增大而增大，在本文中，使用 GPT2-large[55]模型作为计算 MAUVE 指标采用的自回归语言模型，MAUVE 指标用于评价开放问答和术语解释任务。

在具体计算过程中，HuggingFace 开源的 evaluate[56]库为指标计算提供了较为便捷的路径和方法，只需将模型、相关指标计算代码准备完毕，即可通过 evaluate 库加载相关指标进行计算和输出。为方便最终比较，本文将所有指标得分按照百分制进行统一，得到最终各任务的具体分值。

4 评测结果与分析/Evaluation results and analysis

本次评测采用的深度学习框架为 pytorch-2.0.1，模型调用基于 transformers-4.30.1 完成，在硬件环境方面，采用单卡 NVIDIA RTX A6000 48GB 进行模型推理，NVIDIA 驱动版本为 535.146.02，CUDA 版本为 12.2。在实验过程中，考虑到本文涉及任务参考文本长度均不超过 512，为保证模型不会输出过长内容，设置生成文本最大长度为 512，模型输出涉及参数如表 4 所示，表中未列出参数使用默认值。

表 4 模型输出参数

Table 4 Parameters of Model Output

参数名称	参数含义	参数值
max_length	生成序列最大长度	512
min_length	生成序列最小长度	None
do_sample	是否开启采样	False
no_repear_ngram_size	控制重复词生成	0
top_k	保留多少个最高概率词作为候选	40
top_p	已知生成各词总概率为 1，若 top_p 小于 1，则从高到低累加至 top_p，取其中词作为候选	0.9
temperature	控制 softmax 输出的差距	0.2

4.1 领域基础知识

4.1.1 单项选择

单项选择是当前大模型评测的常用任务，考虑到基座模型的指令跟随能力较弱，为提高评测的准确性，本文对基座模型输出进行了后处理，将模型输出不规范但正确的情况修改为了正确选项。根据模型输出的具体情况来看，大多数基座模型难以按照要求输出选项，有些模型只是在单纯的复述题目，而对话模型大多可以输出较为正常的结果，虽然也会夹杂解释内容，但效果远优于基座模型。在各模型中，Baichuan2-7B-Chat 和 InternLM-7B-Chat 模型在输出内容的规范性方面表现优越，在评测的 100 条题目中，做到了完全按照指令，只输出选项。各模型在单项选择任务中的具体表现如表 5 所示。

表 5 单项选择性能指标

Table 5 Indicators of Single choices

模型名称	模型类型	模型得分(Accuracy)
Atom-7B	基座模型	3
Baichuan-7B		10
Baichuan2-7B		16
Chinese-Alpaca-7B		18
InternLM-7B		12
Qwen-7B		**25**
Atom-7B-Chat	对话模型	13
Baichuan2-7B-Chat		**71**
ChatGLM-6B		35
ChatGLM2-6B		52
InternLM-7B-Chat		60
Phoenix-Inst-Chat-7B		38
Qwen-7B-Chat		61

可以看到，大多数对话模型表现出了相当优越的性能，Baichuan2-7B-Chat 模型在对话模型中获得了最高的准确率，而 Qwen-7B 模型在基座模型中获得了最高的准确率。值得注意的是，Atom-7B-Chat 模型在单项选择任务中表现劣于一些基座模型，经过检查模型输出内容，发现 Atom-7B-Chat 模型在单项选择任务中的输出内容经常为空，经过多次重复实验，仍然无法避免，可能是因为 Atom-7B-Chat 模型在使用对话模型进行指令微调过程中所使用的数据质量问题导致。

4.1.2 术语解释与开放问答

术语解释与开放问答是评价大模型领域知识和文本生成能力的任务，相较于单项选择，术语解释和开放问答任务对于输出结果没有格式要求，也无需对输出结果进行后处理，但对于开放性生成任务的评价更具挑战。各模型在术语解释与开发问答任务中的具体表现如表 6 所示。

表 6 名词解释、开放问答得分

Table 6 Indicators of Terminology Definition and Open Q&A

模型名称	模型类型	名词解释		开放问答	
		ROUGE-L	MAUVE	ROUGE-L	MAUVE
Atom-7B	基座模型	19.2	4.96	3.46	1.02
Baichuan-7B		23.37	5.66	8.12	6.85
Baichuan2-7B		**31.19**	**29.52**	**15.06**	6.91
Chinese-Alpaca-7B		16.19	23.04	4.92	4.07
InternLM-7B		27.62	7.39	7.92	6.91
Qwen-7B		30.14	20.31	9.48	**8.18**
Atom-7B-Chat	对话模型	33.34	18.86	5.09	10.57
Baichuan2-7B-Chat		**36.25**	39.91	14.19	15.14
ChatGLM-6B		32.38	34.62	16.17	9.11
ChatGLM2-6B		35.08	31.78	14.78	11.24
InternLM-7B-Chat		32.10	**46.75**	**21.91**	12.03
Phoenix-Inst-Chat-7B		26.86	42.33	17.69	7.57
Qwen-7B-Chat		30.48	28.98	21.86	**15.39**

整体来看，开放问答得分明显低于名词解释，而且，相较于单项选择，这两个任务中基座模型与对话模型的差异较小。究其缘由，当前的生成式模型所表现出来的对话能力本质上是对于文本的续写，而名词解释、开放问答任务的提问方式、指令与日常对话有很大的相似之处，很多模型在预训练阶段采用的数据就会包含一部分问答内容，这也使得基座模型也能表现出不错的对话、问答能力。开放问答之所以分值较低，是因为开放问答用词更为多变，答案多样性更强，因此会影响到使用标准答案进行比对评测的得分计算。

在基座模型中，Baichuan2-7B 模型取得了最佳得分，而在对话模型中，Baichuan2-7B-Chat 模型和 Qwen-7B-Chat 模型分别取得了名词解释和开放问答的最佳得分。在开放问答任务中， Chinese-Alpaca-7B 模型多次输出内容为空，经过反复实验仍然无法避免，而 Atom-7B 系列模型虽然没有输出内容为空的情况，但是出现输出内容非正常语言，例如重复的数字、标点等，考虑到模型评测的实际应用性，并未对这些内容进行后处理而直接计算指标，这使得 Chinese-Alpaca-7B 和 Atom-7B 系列模型指标明显落后于其他模型。

4.2 学术文本

4.2.1 学术文本学科与语步分类

对于分类任务，即便给定了一定量的示例样本，基座模型仍然难以按照正确的格式输出，与单项选择任务类似，本文也对分类任务的基座模型输出进行了后处理，使得评测指标能够更好地反应模型实际性能。经过后处理，得到各模型分类任务具体得分如表 7 所示。

表 7 分类任务得分

Table 7 Indicators of Categories Tasks

模型名称	模型类型	学科分类		语步分类	
		Accuracy	F-1	Accuracy	F-1
Atom-7B	基座模型	4.67	3.3	22	10.44
Baichuan-7B		11.33	13.20	16.5	5.95
Baichuan2-7B		**58.33**	**61.79**	20	6.72
Chinese-Alpaca-7B		40	46.20	20	6.9
InternLM-7B		20	15.28	20	6.69
Qwen-7B		51.67	56.06	**24**	**15.58**
Atom-7B-Chat	对话模型	27	34.46	20	6.93
Baichuan2-7B-Chat		33	26.93	27	18.33
ChatGLM-6B		5.67	2.2	23.5	18.45
ChatGLM2-6B		43	42.87	22.5	14.55
InternLM-7B-Chat		**66.67**	**68.58**	**46**	**42.32**
Phoenix-Inst-Chat-7B		37.67	32.44	20	6.67
Qwen-7B-Chat		37.67	38.45	44	40.22

总体来看，在基座模型中，Baichuan-7B 与 Qwen-7B 模型取得了最佳得分，对话模型中，InternLM-7B-Chat 模型取得了最佳得分。大部分情况下，语步分类得分要低于学科分类，在观察过程中发现，大多数模型仅输出了 2-3 种语步类别，并未将给出的全部语步类别涵盖。此外，Baichuan2-7B、Qwen-7B 模型在学科分类任务种的性能超过了对应对话模型，这一方面是因为对基座模型输出进行了后处理，另一方面，在经过对话数据微调后，应对学科分类这一非正常对话任务，对话模型性能反而下降了。

4.2.2 学术文本标题生成

摘要文本生成一直是自然语言处理领域的传统任务，针对学术文本来说，利用全文内容生成摘要需要提高模型的最大输出长度，同时需要解决其中图标、公式相关内容的解析，难度较高。为评测大模型在文本组织方面的能力，本文采用标题生成任务来代替摘要生成，对

大模型文本归纳能力进行评价，虽然模型仍然会输出一些无关内容，但考虑到本文所采用评测指标 ROUGE 着重计算召回率，所以未对各模型生成文本进行进一步处理，各模型标题生成具体得分如表 8 所示。

表 8 标题生成得分

Table 8 Indicators of Title Generation

模型名称	模型类型	标题生成		
		ROUGE-1	ROUGE-2	ROUGE-L
Atom-7B	基座模型	22.82	10.51	19.17
Baichuan-7B		19.39	9.18	14.21
Baichuan2-7B		29.21	23.85	29.91
Chinese-Alpaca-7B		20.35	10.94	11.53
InternLM-7B		21.97	10.42	13.04
Qwen-7B		**53.23**	**39.02**	**45.94**
Atom-7B-Chat	对话模型	15.88	7.01	9.88
Baichuan2-7B-Chat		49.59	35.55	42.56
ChatGLM-6B		45.80	30.42	36.28
ChatGLM2-6B		48.31	33.69	40.79
InternLM-7B-Chat		52.08	37.47	44.66
Phoenix-Inst-Chat-7B		28.86	16.67	18.67
Qwen-7B-Chat		**52.69**	**37.89**	**45.28**

在对话模型与基座模型中，分别由 Qwen-7B-Chat 和 Qwen-7B 模型取得最佳得分，而 Qwen-7B 模型的性能甚至略微超过了 Qwen-7B-Chat 模型。结合 Qwen-7B 模型在前几个任务的表现，可以肯定的是 Qwen 模型在预训练阶段就加入了对话数据进行训练，也确实提升了基座模型的性能，使得 Qwen 模型的对话能力明显优于其他基座模型。

4.2.3 学术文本翻译

机器翻译也是自然语言处理领域的传统任务，在大语言模型不断发展的当下，大模型在机器翻译方面表现出的优越性能改变了机器翻译的研究范式，机器翻译由之前的 Encoder + Decoder 架构逐渐向 Decoder-only 架构转变。本文通过包含学术术语的中英平行语料，对大模型对于人文社科领域学术文本的翻译能力进行评测，各模型文本翻译具体得分如表 9 所示。

表 9 文本翻译得分

Table 9 Indicators of Translation

模型名称	模型类型	英-中		中-英	
		BLEU	chrF	BLEU	chrF
Atom-7B	基座模型	2.23	3.75	3.50	8.17
Baichuan-7B		16.33	4.84	13.88	25.17
Baichuan2-7B		20.21	19.06	24.86	30.69
Chinese-Alpaca-7B		15.17	15.35	9.77	9.24
InternLM-7B		19.30	27.36	16.69	24.72
Qwen-7B		**32.13**	**28.20**	**60.84**	**53.75**
Atom-7B-Chat	对话模型	4.24	4.57	15.51	17.44
Baichuan2-7B-Chat		**32.54**	**28.36**	**60.75**	**55.51**
ChatGLM-6B		27.17	23.88	55.53	48.57
ChatGLM2-6B		27.14	23.58	55.11	48.34

InternLM-7B-Chat	25.77	23.13	51.97	46.44
Phoenix-Inst-Chat-7B	23.67	15.88	24.90	25.76
Qwen-7B-Chat	31.43	27.50	60.41	54.13

文本翻译中文互译任务中，Qwen-7B 与 Baichuan2-7B-Chat 模型分别获得基座模型与对话模型中的最佳得分。对比两个子任务上的不同模型得分，可以看到，中-英翻译得分显著高于英-中翻译，结合评测数据和模型输出内容，本文认为原因有二。一方面，文本翻译指标的计算主要基于生成文本中的 n-gram 数量，而英文文本长度(词数)要低于中文文本长度(字数)，因此在指标计算上，英文结果会有更高的得分。另一方面，经过对各模型输出文本的检查，发现在英文-中文翻译任务的输出结果中，包含一部分输出结果为英文，这可能是因为模型无法正确理解指令而直接对给出的待翻译英文文本进行续写导致。

在各项任务上，对于绝大多数模型来说，对话模型有着更加优越的性能，也有一些模型表现例外，例如 Qwen 无论是基座模型还是对话模型在文本翻译上均表现出了优越的性能，且基座模型性能略优于对话模型，而 Atom 对话模型在标题生成任务上表现很差，甚至弱于其基座模型。此外，对于学科分类任务，经过后处理的 Baichuan2、Qwen 基座模型性能明显优于对话模型，这也表明这些模型在人文社科领域有一定的知识储备，而对话模型也并不能完全按照指令进行输出。总的来说，与本文开始的假设基本一致，即便是在人文社科这一垂直领域，对于同一个模型来说，在不对结果进行后处理的情况下，对话模型所表现出来的性能几乎是完全优于基座模型的。

综上所述，本文基于 7 个任务对 6 个基座大模型和 7 个对话大模型进行了评测，为得到一个更加直观的各模型性能对比分数，考虑到不同任务指标在数值上差异性较大，本文使用各模型在不同任务上的排名计算各模型的最终得分。具体来说，对于基座模型，某个任务排名第一模型得六分，以此递减至最后一名，最终得到模型的综合评价分数如表 10 所示。

表 10 各模型最终得分

Table 10 Final scores of Models

模型名称	模型类型	领域知识得分	学术文本得分	综合得分
Atom-7B	基座模型	3	12	15
Baichuan-7B		8	10	18
Baichuan2-7B		16	23	39
Chinese-Alpaca-7B		11	14	25
InternLM-7B		9	17	26
Qwen-7B		16	29	**45**
Atom-7B-Chat	对话模型	3	7	10
Baichuan2-7B-Chat		19	27	**46**
ChatGLM-6B		10	18	28
ChatGLM2-6B		12	21	33
InternLM-7B-Chat		17	26	43
Phoenix-Inst-Chat-7B		7	11	18
Qwen-7B-Chat		16	30	**46**

Qwen 模型无论是基座模型还是对话模型均获得了最高的得分，而且，Qwen-7B 基座模型在标题生成、文本翻译方面相较于其他模型有着断档式的领先，甚至不弱于一些对话模型。结合各任务 Qwen-7B 基座模型表现，Qwen-7B 模型在预训练阶段就以某种形式加入了对话数据，这使得其在常见的自然语言处理任务中表现出了非常优越的性能。除此之外，在单项选择任务中，Baichuan2-7B-Chat 和 InternLM-7B-Chat 两个模型几乎完全按照指令，只输出

选项，这对于生成式语言模型是非常艰巨的任务，可能是对话模型构建过程中加入的特殊指令数据带来的效果。

总的来说，在各项任务上，对话模型基本表现出了优于基座模型的性能，指令微调作为大模型构建中承前启后的一环，其更多是将预训练阶段注入到模型的数据、知识引导出来，使模型能够更好地理解自然语言。由此可见，当前垂直领域模型的构建，一方面要注重领域数据集的构建，保证有充足的领域知识对模型进行增强，另一方面也要注重模型自身的对话能力，使模型能够充分发挥其能力。

5 总结与展望/Conclusions and prospects

本文针对人文社科领域知识和学术文本构建了一系列大模型评测任务，旨在为人文社科领域研究人员提供大模型研究的参考，助力人文社科研究与人工智能技术的交叉融合。通过对 13 个基座、对话模型进行评测，评测结果显示，在人文社科领域，Qwen 系列模型有着最为优越的性能，Baichuan2 系列模型紧随其后，InternLM 模型位列第三，Atom 模型表现最差。同时，通过对模型输出内容的分析，本文还发现，影响基座模型得分的一个重要因素，即是模型在输出过程中容易输出过多无效内容，无法理解指令含义。由此可见，对于蕴含庞大知识量的大模型来说，教会其理解指令是提升大模型性能的重要途经，这也是当前一个合格的大模型必须经历的三个阶段之二。要训练一个合格的大模型，需要经历三个阶段：海量文本预训练，将大量文本数据注入到大模型中，供大模型学习，使得大模型拥有足够的知识储备以应对不同领域的问题；对话数据指令微调，这一阶段通过多任务指令学习，让大模型学会如何使用预训练阶段注入的知识，使模型具备更加多样的能力；人类反馈强化学习，将大模型价值观与人类对齐，使其输出更符合人类偏好。当前绝大多数开源对话模型为第二阶段的产物，根据本文的评测结果也可以看到，很多时候，大模型并非对领域知识有所缺失，而是对话能力不足。

当然，一些表现异常优异的模型也带来了当前大模型评测的另一个挑战，由于当前模型训练数据的封闭性，无法判断模型表现出的优越性能是其本身能力还是训练数据的特殊化导致。这就带来了两个方面亟待解决的问题，模型可解释性和更为全面准确的大模型评价体系，神经网络发展至今，可解释性仍然是一个难以克服的问题，而大模型的出现进一步丰富了该问题，大模型“涌现”能力的根源、如何对大模型进行有效解释、大模型安全等问题成为了当前模型可解释性研究的又一路径[57]。本文虽然基于人文社科领域提出了评价大模型领域能力的评测体系，但仍存在一定的不足之处，具体来说，评测指标有待改进，生成内容很难通过定量指标进行评价，评价方法、指标的扩充丰富是大模型评测能够进一步发展的必要保证。

自 2022 年 ChatGPT 发布以来，经历了一年多的时间，大模型研究逐渐形成了受到普遍认可的研究路径和范式，时至今日，大模型研究需要解决主要有三个问题：算法、算力、数据。算法层面，目前模型架构呈现 Decoder-only 架构为主流其他少数其他框架各自发展的态势，算法方面逐渐趋同化；算力层面，硬件设施导致很多普通的科研团队难以加入到大模型的研究之中，相关研究人员也在探索能够尽可能降低模型训练成本的方法，LoRA[58]在大模型上的应用使得更多科研人员能够加入到这场大模型研究的浪潮之中；数据层面，在人工智能技术壁垒逐渐降低的当下，数据质量成为大模型性能的基础，作为信息资源管理专业的研究人员，我们既要发挥专业优势，做好数据收集、组织、存储、服务等一系列数据工程，在数据构建的全流程中，探索更多路径和方法，也要注重学科之间的交叉融合，将技术手段运用到各个领域，推动大数据时代向大知识、大智能时代发展。

参考文献：

[1] 黄水清, 刘浏, 王东波. 计算人文的发展及展望[J]. 科技情报研究, 2021, 3(4): 1-12. (H

UANG S Q, LIU L, WANG D B. The development and outlook of computing huma nities[J]. Scientific information research, 2021, 3(4): 1-12.)

[2] 丁波涛. 计算社会科学相关概念的比较与辨析[J]. 情报资料工作, 2018(6): 60-67. (DING B T. Comparison and analysis of computational social science related concepts[J]. Information and documentation services, 2018(6): 60-67)

[3] 黄水清, 刘浏, 王东波. 计算人文学科的内涵、体系及机遇[J]. 图书与情报, 2023(1): 1-11+153+145. (HUANG S Q, LIU L, WANG D B. The connotation, system and opportunity of computational humanities [J]. Library & information, 2023(1): 1-11+153+145.)

[4] 王东波, 刘畅, 朱子赫, 等. SikuBERT 与 SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究[J]. 图书馆论坛, 2022, 42(6): 31-43. (WANG D B, LIU C, ZHU Z H, et al. Construction and application of pre-trained models of Siku Quanshu in orientation to digital humanities [J]. Library tribune, 2022, 42(6): 31-43.)

[5] 张卫, 王昊, 邓三鸿, 等. 面向数字人文的古诗文本情感术语抽取与应用研究[J]. 中国图书馆学报, 2021, 47(4): 113-131. (ZHANG W, WANG H, DENG S H, et al. Sentiment term extraction and application of Chinese ancient poetry text for digital humanities [J]. Journal of library science in China, 2021, 47(4): 113-131.)

[6] 张琪, 王东波, 黄水清, 等. 史书多维知识重组与可视化研究——以《史记》为对象[J]. 情报学报, 2022, 41(02): 130-141. (ZHANG Q, WANG D B, HUANG S Q, et al. Multi-dimensional knowledge reorganization and visualization of history books: based on records of the grand historian [J]. Journal of the China society for scientific and technical information, 2022, 41(02): 130-141.)

[7] 喻雪寒, 何琳, 王献琪. 基于机器阅读理解的古文事件抽取研究[J]. 情报学报, 2023, 42(3): 316-326. (YU X H, HE L, WANG X Q. Research on event extraction from ancient books based on machine reading comprehension [J]. Journal of the China society for scientific and technical information, 2023, 42(3): 316-326.)

[8] 张瑞祥, 赵志枭. 人工智能视域下计算法学的概念、探究及趋势[J]. 图书与情报, 2023(1): 39-47. (ZHANG R X, ZHAO Z X. Concepts, explorations and trends of computational jurisprudence from the perspective of artificial intelligence [J]. Library & information, 2023(1): 39-47.)

[9] 梁柱, 沈思, 叶文豪, 等. 基于结构内容特征的裁判文书自动推荐研究[J]. 情报学报, 2022, 41(2): 167-175. (LIANG Z, SHEN S, YE W H, et al. Automatic recommendation of judgment documents based on structural content features[J]. Journal of the China society for scientific and technical information, 2022, 41(2): 167-175.)

[10] WU S, IRSOY O, LU S, et al. Bloomberggpt: a large language model for finance[J]. arXiv preprint arXiv:2303.17564, 2023.

[11] YANG H, LIU X Y, WANG C D. FinGPT: open-source financial large language models[J]. arXiv preprint arXiv:2306.06031, 2023.

[12] XIE Q, HAN W, ZHANG X, et al. PIXIU: a large language model, instruction data and evaluation benchmark for finance[J]. arXiv preprint arXiv:2306.05443 arXiv, 2023.

[13] YANG Y, TANG Y, TAM K Y. InvestLM: a large language model for investment using financial domain instruction tuning[J]. arXiv preprint arXiv:2309.13064 arXiv, 2023.

[14] LawGPT_zh[EB/OL]. [2023-10-11]. https://github.com/LiuHC0428/LAW-GPT.

[15] CUI J X, LI Z J, YAN Y, et al. ChatLaw: open-Source legal large language model

with integrated external knowledge bases[J]. arXiv preprint arXiv:2306.16092 arXiv, 20 23.
[16] Blcuicall/taoli[EB/OL]. [2024-01-21]. https://github.com/blcuicall/taoli
[17] ECNU-ICALK/EduChat[EB/OL]. [2024-01-21]. https://github.com/icalk-nlp/EduChat.
[18] Scutcyr/BianQue[EB/OL]. [2024-01-21]. https://github.com/scutcyr/BianQue.
[19] SCIR-HI/Huatuo-Llama-Med-Chinese[EB/OL]. [2024-01-21]. https://github.com/SCIR-HI/ Huatuo-Llama-Med-Chinese.
[20] Li Y, Ma S, Wang X, et al. EcomGPT: Instruction-tuning large language models with chain-of-task tasks for e-commerce[J]. arXiv preprint arXiv:2308.06966, 2023.
[21] IMOSR/MediaGPT[EB/OL]. [2024-01-21]. https://github.com/IMOSR/MediaGPT.
[22] 生成式人工智能服务管理暂行办法[EB/OL]. [2023-10-10]. https://www.gov.cn/zhengce/z hengceku/202307/content_6891752.htm. (Interim measures for the management of gener ative artificial intelligence services[EB/OL]. [2023-10-10]. https://www.gov.cn/zhengce/z hengceku/202307/content_6891752.htm.)
[23] WANG Z, XIE Q, DING Z, et al. Is ChatGPT a good sentiment analyzer? a prelimin ary study[J]. arXiv preprint arXiv:2304.04339, 2023.
[24] ZHANG W, DENG Y, LIU B, et al. Sentiment analysis in the era of large language models: A reality check[J]. arXiv preprint arXiv:2305.15005, 2023.
[25] PENA A, MORALES A, FIERREZ J, et al. Leveraging large language models for top ic classification in the domain of public affairs[J]. arXiv preprint arXiv:2306.02864, 20 23.
[26] ZHU W, ZHOU H, GAO C, et al. Research development of machine translation and large language model[C]//Proceedings of the 22nd Chinese national conference on com putational linguistics (volume 2: frontier forum). 2023: 30-39.
[27] DAO X Q, LE N B. Investigating the effectiveness of ChatGPT in mathematical reas oning and problem solving: evidence from the Vietnamese national high school gradua tion examination[J]. arXiv preprint arXiv:2306.06331, 2023.
[28] WU Y, JIA F, ZHANG S, et al. An empirical study on challenging math problem sol ving with GPT-4[J]. arXiv preprint arXiv:2306.01337, 2023.
[29] DUONG D, SOLOMON B D. Analysis of large-language model versus human perfor mance for genetics questions[J]. European journal of human genetics, 2023: 1-3.
[30] GILSON A, SAFRANEK C W, HUANG T, et al. How does ChatGPT perform on th e United States medical licensing examination? The implications of large language mo dels for medical education and knowledge assessment[J]. JMIR medical education, 202 3, 9(1): e45312.
[31] ZHAO J, FANG M, SHI Z, et al. CHBias: bias evaluation and mitigation of Chinese conversational language models[J]. arXiv preprint arXiv:2305.11262, 2023.
[32] PARRISH A, CHEN A, NANGIA N, et al. BBQ: a hand-built bias benchmark for qu estion answering[C]//Findings of the association for computational linguistics: ACL 202 2. Dublin, Ireland: Association for computational linguistics, 2022: 2086-2105.
[33] ZHANG W, ALJUNIED SVM, GAO C, et al. M3Exam: a multilingual, multimodal, multilevel benchmark for examining large language models[J]. arXiv preprint arXiv:230 6.05179, 2023.
[34] HUANG Y, BAI Y, ZHU Z, et al. C-eval: a multi-level multi-discipline chinese evalu

ation suite for foundation models[J]. arXiv preprint arXiv:2305.08322, 2023.

[35]XU L, LI A, ZHU L, et al. SuperCLUE: a comprehensive Chinese large language model benchmark[J]. arXiv preprint arXiv:2307.15020, 2023.

[36] Open LLM leaderboard - a HuggingFace space by HuggingFaceH4[EB/OL]. [2023-10-10]. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

[37] Chinese-llm-benchmark[EB/OL]. (2023-10-10)[2023-10-10]. https://github.com/jeinlee1991/chinese-llm-benchmark.

[38] FlagAlpha/Llama2-Chinese[EB/OL]. [2023-10-07]. https://github.com/FlagAlpha/Llama2-Chinese.

[39] Baichuan-Inc/Baichuan-7B: A large-scale 7B pretraining language model developed by BaiChuan-Inc.[EB/OL]. [2023-10-07]. https://github.com/baichuan-inc/Baichuan-7B.

[40] Baichuan-Inc/Baichuan2: A series of large language models developed by Baichuan-Inc [EB/OL]. [2023-10-07]. https://github.com/baichuan-inc/Baichuan2.

[41] Ymcui/Chinese-LLaMA-Alpaca[EB/OL]. [2023-10-07]. https://github.com/ymcui/Chinese-LLaMA-Alpaca.

[42] InternLM/InternLM[EB/OL]. [2023-10-07]. https://github.com/InternLM/InternLM.

[43] QwenLM/Qwen[EB/OL]. [2023-10-07]. https://github.com/QwenLM/Qwen.

[44] THUDM/ChatGLM-6B[EB/OL]. [2023-10-07]. https://github.com/THUDM/ChatGLM-6B.

[45] THUDM/ChatGLM2-6B[EB/OL]. [2023-10-07]. https://github.com/THUDM/ChatGLM2-6B.

[46] FreedomIntelligence/LLMZoo[EB/OL]. [2023-10-07]. https://github.com/FreedomIntelligence/LLMZoo.

[47]魏向清. 人文社科汉英术语知识库构建探索：理念与方法[M]. 1 版. 南京: 南京大学出版社, 2022. (WEI X. Exploring the construction of Chinese-English terminology knowledge base in humanities and social sciences: concepts and methods [M].1st ed. Nanjing: Nanjing University Press,2022.)

[48] ChatGPT prompt engineering for developers[EB/OL]. [2023-10-07]. https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/.

[49] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. ACM computing surveys, 2023, 55(9): 1-35.

[50] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the association for computational linguistics. 2002: 311-318.

[51] LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]//Text summarization branches out. Barcelona, Spain: Association for computational linguistics, 2004: 74-81.

[52] POPOVIC M. chrF: character n-gram F-score for automatic MT evaluation[C]//Proceedings of the tenth workshop on statistical machine translation. 2015: 392-395.

[53] POST M. A call for clarity in reporting BLEU scores[C]//Proceedings of the third conference on machine translation: research papers. Brussels, Belgium: Association for computational linguistics, 2018: 186-191.

[54] PILLUTLA K, SWAYAMDIPTA S, ZELLERS R, et al. Mauve: measuring the gap between neural text and human text using divergence frontiers[J]. Advances in neural i

nformation processing systems, 2021, 34: 4816-4828.
[55] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
[56] Evaluate metric[EB/OL]. [2023-10-08]. https://huggingface.co/evaluate-metric.
[57] ZHAO H, CHEN H, YANG F, et al. Explainability for large language models: a survey[J]. arXiv preprint arXiv:2309.01029, 2023.
[58] HU E J, SHEN Y, WALLIS P, et al. Lora: low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.

作者贡献说明：赵志枭：完成实验流程，论文撰写；胡蝶：实验流程完善，论文撰写；刘畅：实验流程完善，论文修改；沈思：论文修改；王东波：论文整体框架及研究思路制定。

Performance Evaluation of Chinese Universal Large Model in the Field of Humanities and Social Sciences

Zhao Zhixiao[1,2] Hu Die[1,2] Liu Chang[1,2] Shen Si[3] Wang Dongbo[1,2]
[1]College of Information Management, Nanjing Agricultural University, Nanjing 210095
[2]Research Center of Humanities and Social Computing, Nanjing Agricultural University, Nanjing 210095
[3]School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094

Abstract: [Purpose/Significance] This paper starts from the field of humanities and social sciences, and compares the model performance of humanities and social sciences from the aspects of basic knowledge and academic texts of humanities and social sciences. It aims to provide a systematic large language model evaluation benchmark for the field of humanities and social sciences for the reference of researchers in humanities and social sciences related fields. **[Methods/Processes]** Seven evaluation tasks related to the field of humanities and social sciences were designed and corresponding indicators were selected. On this basis, the current open-source and high-performance general-purpose domain Chinese large language models were selected to complete the domain-specific tasks in the form of questions and answers by invoking the local models, and their performance in the field of humanities and social sciences was quantitatively evaluated by selecting relevant indicators. **[Results/Conclusions]** The evaluation results show that among the open-source models selected in this paper, Qwen has the best performance, followed by Baichuan2, InternLM, and Atom is the worst performer in both the base model and the dialog model; moreover, in most cases, the dialog model shows more superior performance compared to the base model.

Keywords: Humanities and Social Science Large Model Evaluation Domain Knowledge Academic Texts